KEEP THIS COPY FOR REPRODUCTION PURPOSES

AD-A238 342

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

(2)

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE: 5/1/91

3. REPORT TYPE AND DATES COVERED: Final 1 May 86 - 31 December 90

4. TITLE AND SUBTITLE: Data Analysis

5. FUNDING NUMBERS: DAAL03-86-K-0073

6. AUTHOR(S): John W. Tukey

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES):
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES):
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER: ARO 23360.18-MA

11. SUPPLEMENTARY NOTES:
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT:
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The papers that follow sketch the content of 60 publications, of which 7 are books. It also covers 8 technical reports and 13 internal working papers, as well as considerable amounts of unreported and unpublished material.

Some of the more innovative contributions brought to full exposition include:

1) A volume on Configural Polysampling, so far the only direct approach to robustness in finite-sized samples.

2) A volume on Exploratory Analysis of Variance, taking an approach which is both novel and effective.

Other important innovations include:

3) New light on multiple-comparison problems, especially as they arise in the analysis of variance.

4) Use of simple regressions (on an orthogonal space) as a means of composite building.

5) Development of general techniques of shape comparison

6) Progress on the "separations problem".

7) Discussion of how resampling methods (jackknife, bootstrap) should be applied to problems where blocking (in the design of experiment sense) is essential.

8) Use of limited lateral randomization in visualizing distributions.

9) Introduction of novel, more effective measures of urbanization.

10) Development of new, apparently promising approaches to clustering.

A-1

14. SUBJECT TERMS:
Data analysis, Statistics, Robustness, Analysis of variance, Mutliple comparisons, Shape comparison, Resampling with blocking, Limited lateral randomization

15. NUMBER OF PAGES: 17

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

91-05321

91

Data Analysis

by

John W. Tukey

FINAL REPORT
May 1, 1986 — December 31, 1990
DAAL03-86-K-0073

supported by the

U. S. Army Research Office (Durham)

TABLE OF CONTENTS

Executive Summary

The papers that follow sketch the content of 62 publications, of which 7 are books. It also covers 12 technical reports and 13 internal working papers, as well as considerable amounts of unreported and unpublished material.

Some of the more innovative contributions brought to full exposition include:

1) A volume on Configural Polysampling, so far the only direct approach to robustness in finite-sized samples (Tukey 1991c).

2) A volume on Exploratory Analysis of Variance, taking an approach which is both novel and effective (Tukey 1991h).

Other important innovations include:

3) New light on multiple-comparison problems, especially as they arise in the analysis of variance (Tukey 1991n).

4) Use of simple regressions on orthogonal space as a means of composite building where the data is not strong enough for conventional multivariate techniques (Tukey 1991o).

5) Development of general techniques of shape comparison that take advantage of such available machinery as weighted least squares and conventional multivariate analysis (Goodall, see Section 15, below).

6) Progress on the "separations problem" where we ask if a batch of numbers is better thought of as two (or more) subbatches (Technical Reports 293, 298).

7) Discussion of how resampling methods (jackknife, bootstrap) should be applied to problems where blocking (in the design of experiment sense) is essential (Technical Report 292).

8) Use of limited lateral randomization in visualizing distributions involving many (100 to 10,000) points (Tukey and Tukey 1991Se).

9) Introduction of novel, more effective measures of urbanization (Kafadar and Tukey 1990Sa).

10) Development of new, apparently promising approaches to clustering (Hansen & Tukey 1990Sb).

John W. Tukey

John W. Tukey
Princeton, 2 May 1991

1. Personnel 1986 — 1990

Faculty

Colin Goodall 1986-87
John W. Tukey 1986-90

Visiting Faculty (short term)

Thu Hoang April-June 1987, February 1988, March 1989
Catherine Marsh September 1988
Karen Kafadar December 1989
Kaye Basford January-February 1990

Graduate Students

Ha Nguyen 1986 (Ph D. 1986)
Katherine M. Hansen (Ph.D. 1989)

Undergraduate Student

Bill Frack August-September 1987

Research Assistant

E. Olszewski 1986-90

2. Publications 1986 — 1990

Books, chapters in books:

Tukey, John W. (1986c). *The Collected Works of John W. Tukey, Volume III: Philosophy and Principles of Data Analysis, 1949 - 1964.* (L. V. Jones, ed.) Wadsworth Advanced Books & Software, Monterey, CA.

Tukey, John W. (1986d). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. *The Collected Works of John W. Tukey, Volume III: Philosophy and Principles of Data Analysis, 1949-1964. 187-389.* Wadsworth Advanced Books & Software, Monterey, CA.

Tukey, John W. (1986e). *The Collected Works of John W. Tukey, Volume IV: Philosophy and Principles of Data Analysis, 1965 - 1986.* (L. V. Jones, ed.) Wadsworth Advanced Books & Software, Monterey, CA.

Tukey, John W. (1986f). "What have statisticians been forgetting? *The Collected Works of John W. Tukey, Volume IV: Philosophy and Principles of Data Analysis, 1965-1986.* 587-599. Wadsworth Advanced Books & Software, Monterey, CA.

NOTE: Letters used with years on John Tukey's papers correspond to bibliographies in all volumes of his collected papers.

Tukey, John W. (1986g). Choosing techniques for the analysis of data. Wadsworth Advanced Books & Software, Monterey, CA. *The Collected Works of John W. Tukey. Volume IV: Philosophy and Principles of Data Analysis, 1965-1986.* 869-874. Wadsworth Advanced Books & Software, Monterey, CA.

Tukey, John W. (1986h). Do derivations come from heaven? *The Collected Works of John W. Tukey. Volume IV: Philosophy and Principles of Data Analysis, 1965-1986.* 875-880. Wadsworth Advanced Books & Software, Monterey, CA.

Tukey, John W. (1988a). *The Collected Works of John W. Tukey, Volume V: Graphics, 1965 - 1985.* (W. S. Cleveland, ed.) Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1988b). Data analysis and statistics: Techniques and approaches. *The Collected Works of John W. Tukey. Volume V: Graphics, 1965-1985.* 1-22. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1988c). Notch plots for counted rates. *The Collected Works of John W. Tukey. Volume V: Graphics, 1965-1985.* 79-92. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1988d). Control and stash philosophy for two-handed, flexible, and immediate control of graphic display. *The Collected Works of John W. Tukey. Volume V: Graphics, 1965-1985.* 329-382. Wadsworth Advanced Books & Software, Pacific Grove, CA. [Also in *Dynamic Graphics for Statistics.* (W. S. Cleveland and M. E. McGill, eds) 133-178. Wadsworth Advanced Books & Software, Pacific Grove, CA.]

Tukey, John W. (1988e). Thoughts on the evolution of dynamic graphics for data-modification display. *The Collected Works of John W. Tukey. Volume V: Graphics, 1965-1985.* 383-401. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1990a). *The Collected Works of John W. Tukey, Volume VI: More Mathematical, 1938 - 1984.* (C. L. Mallows, ed.) Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1990b). The finite case of the "Problem of the Nile." *The Collected Works of John W. Tukey. Volume VI: More Mathematical, 1938-1984.* 35-40. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1990c). The present state of fiducial probability. *The Collected Works of John W. Tukey. Volume VI: More Mathematical, 1938-1984.* 55-118. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1990d). Handouts for the Wald lectures 1958. *The Collected Works of John W. Tukey. Volume VI: More Mathematical, 1938-1984.* 119-148. Wadsworth Advanced Books & Software, Pacific Grove, CA.

Tukey, John W. (1990e). Souvenir sheets for "the criticism of transformations." *The Collected Works of John W. Tukey. Volume VI: More Mathematical, 1938-1984.* 157-165. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990f). The practical relationship between the common transformations of percentages or fractions and of amounts. *The Collected Works of John W. Tukey. Volume VI: More Mathematical, 1938-1984.* 211-219. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990g). Introduction to guided re-expression. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 221-236. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990h). Determination of linear relations between systematic parts. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 317-329. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990i). Standard confidence points. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 331-365. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990j). Some general principles and approximations for sequentially studentized procedures. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 367-371. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990k). One degree of freedom or several? Parsimony in detection of an effect. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 407-419. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990l). A note on least squares and unbiased estimates. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 421-427. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990m). Better coordinates for combinations of order statistics. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 429-433. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990n). Souvenir sheet for "Use of control classifications: Adjustment for inadequacy of broad classes." *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 435-443. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990o). NkNS forests. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 553-556. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990p). Some further multivariate suggestions. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 557-583. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1990q). Steps toward a universal univariate distribution analyzer. *The Collected Works of John W. Tukey , Volume VI: More Mathematical, 1938-1984.* 585-590. Wadsworth Advanced Books & Software, Pacific Grove. CA.

Tukey, John W. (1991a). *The Collected Works of John W. Tukey, Volume VII: Factorial & Anova.* (David R. Cox, ed.) Wadsworth Advanced Books & Software, Pacific Grove, CA. To appear.

Tukey, John W. (and Morgenthaler, Stephan, as co-editors) (1991c). *Configural Polysampling: A Route to Practical Robustness.* Wiley & Sons, Inc., New York.

Tukey, John W. (and Morgenthaler, Stephan) (1991d). Chapter 1: Background. *Configural Polysampling: A Route to Practical Robustness.* (S. Morgenthaler, and J. W. Tukey, eds.) 1-8. John Wiley & Sons, Inc., New York,

Tukey, John W. (and Morgenthaler, Stephan) (1991e). Chapter 3: Key Ideas and Outline. *Configural Polysampling: A Route to Practical Robustness.* (S. Morgenthaler, and J. W. Tukey, eds.) 21-36. John Wiley & Sons, Inc., New York.

Tukey, John W. (and Morgenthaler, Stephan) (1991f). Chapter 7: Point Estimation of Location: Technical Choices. *Configural Polysampling: A Route to Practical Robustness.* (S. Morgenthaler, and J. W. Tukey, eds.) 87-108. John Wiley & Sons, Inc., New York.

Tukey, John W. (1991g). Chapter 12: Appendix on a Circular Approach to Cubature. *Configural Polysampling: A Route to Practical Robustness.* (S. Morgenthaler, and J. W. Tukey, eds.) 213-221. John Wiley & Sons, Inc., New York.

Tukey, John W. (and Hoaglin, David C., Mosteller, Frederick as co-editors) (1991h). *Fundamentals of Exploratory Analysis of Variance.* John Wiley & Sons, Inc., New York. To appear.

Tukey, John W. (with Mosteller, Frederick and Parunak, Anita) (1991i). Mean squares, F. tests, and estimates of variance. *Fundamentals of Exploratory Analysis of Variance.* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds.). John Wiley & Sons, Inc. To appear.

Tukey, John W. (with Hoaglin, David C.) (1991j). Qualitative and quantitative confidence. *Fundamentals of Exploratory Analysis of Variance.* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds.). John Wiley & Sons, Inc. To appear.

Tukey, John W. (with Mosteller, Frederick and Youtz, Cleo S.) (1991k). Assessing changes. *Fundamentals of Exploratory Analysis of Variance.* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds.). John Wiley & Sons, Inc. To appear.

Tukey, John W. (with Mosteller, Frederick) (1991l). Purposes of analyzing data that comes in a form inviting us to apply tools from the analysis of variance. *Fundamentals of Exploratory Analysis of Variance.* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds.). John Wiley & Sons, Inc. To appear.

Tukey, John W. (with Hoaglin, David C., and Mosteller, Frederick) (1991m). Concepts and examples in analysis of variance (ANOVA). *Fundamentals of Exploratory Analysis of Variance.* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.). John Wiley & Sons, Inc. To appear.

Papers:

Goodall, Colin, R., Stoughton, C., and Easton, G. (1986). "Delineation plots for bivariate data: an outline of the method," *Proceedings of the 1986 American Statistical Association* (Section on Statistical Graphics), 81-85.

Goodall, C. R. (1986). Comment on F. L. Bookstein: Size and shape spaces for landmark data. *Statistical Science*, **1:** 191--242.

Goodall, Colin, R. (1987). Review: Modern Multivariate Statistical Analysis: A Graduate Course and Handbook, authors Siotani, M., Hayakawa, T., and Fujikoshi, Y. American Science Press, Columbus, OH *Technometrics*, Vol. **29:** 242-243.

Goodall, C. R. (1987). Statistical analysis of biomedical images: techniques for shape comparisons. Discussion. *American Statistical Association: 1987 Proceedings of the Statistical Graphics Section*, 28-30.

Goodall, C. R., and Thoma, H. M. (1987). Interpolation of multivariate data. *American Statistical Association: 1987 Proceedings of the Statistical Graphics Section*, 64-67.

Goodall, Colin, R. and Bose, A. (1987). Models and procrustes methods for the analysis of shape difference. *Proceedings of the 19th Symposium on the Interface between Computer Science and Statistics*, 86-92.

Goodall, Colin R. and Hansen, Katherine M. (1988). Resistant smoothing of high-dimensional data. *1988 Proceedings of the Statistical Graphics Section, American Statistical Association*, 12-21.

Kafadar, Karen, and Tukey, John W. (1988h). "A bidec *t* table," *Journal of the American Statistical Association*, Vol. **83:** 532-539.

Morgenthaler, Stephan, and Tukey, John W. (1990r). "The next future of data analysis," *Data Analysis, Learning Symbolic and Numeric Knowledge*, (Edwin Diday, ed.), Nova Science Publishers, New York.

Tukey, John W. (1986k). "The interface with computing: in the small or in the large," *Computer Science and Statistics: proceedings of the 18th Symposium on the Interface*. (T. J. Boardman, ed.) 3-7. American Statistical Assoc., Washington, D. C.

Tukey, John W. (1986u). Discussion of paper by D. R. Brillinger [The natural variability of vital rates and associated statistics], *Biometrics* **42:** 729-732.

Tukey, John W. (1987k). Comment on paper by R. A. Becker, W. S. Cleveland and A. R. Wilks [Dynamic graphics for data analysis], *Statistical Science*, **2:** 383-385.

Tukey, John W. (1989w). "SPES in the years ahead," *Proc. of Amer. Statist. Assoc. Sesquicentennial 1988-1989 meetings*, Washington, D. C.

Tukey, John W. (1991n). "The philosophy of multiple comparisons," (1989 Miller Lecture presented at Stanford University), *Statistical Science*, **6:** No. 1, 98-116.

Tukey, John W. (1991o). "Use of many covariates in clinical trials," *International Statistical Review, Vol.* **59:** No. 2, August. To appear.

Tukey, John W. (1991p). "Exbrids: Nearly symmetrizing re-expressions for experimentally distributed quantities," *Essays in Statistics: In Honour of G. S. Watson*, (K. V. Mardia, ed.), John Wiley & Sons, Ltd., Sussex, England. To appear.

Tukey, John W. (1991q). "Consumer datesware," *Directions in Robust Statistics and Diagnostics, Part II, IMA Volumes in Mathematics and its Applications* **34.** (Werner Stahel and Sanford Weisberg, eds.) Springer-Verlag. 297-308.

Goodall, C. R. and De Veaux, R. D. (1991U). Final downsweeping: the use of Paull's rule for aggregation. to appear in *The analysis of Variance, Vol. II* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), John Wiley & Sons.

Submitted for publication:

Cohen, Michael, Dalal, Siddartha R., and Tukey, John W. (1991Sa). *Robust, smoothly - heterogeneous* variance regression," to appear in *Applied Statistics.*

Hansen, Katherine, and Tukey, John W. (1990Sb). "Tuning a major part of a clustering algorithm," *International Statistical Review.* (Revised, to be resubmitted.)

Hoaglin, David C., and Tukey, John W. (1989Sc). "Empirical bounds for quantile-based estimates of g in the g-and-h distributions," *Technometrics.* (In revision).

Kafadar, Karen, and Tukey, John W. (1990Sd). "An approach to U. S. cancer death rates involving *urbanization and geographic contiguity:* 1. *A simple adjustment for* urbanization," *International Statistical Review.* To appear.

Tukey, John W., and Tukey, Paul A. (1990Se). "Strips displaying empirical distributions: I: Textured dot strips,", (prepared in part in connection with research at Princeton *University sponsored by the Army Research Office* (Durham) DAAL03-86-K-0073 and DAAL03-88-K-0045.) Submitted to Journal American of the Statistical Association.

Papers delivered and in preparation:

Goodall, Colin R. (1987a). Multivariate procrustes techniques. Symposium on Shape Theory, Princeton university, May 1987.

Goodall, Colin R. (1987b). Mathematical phylotaxis -- a review. Invited talk, XIV International Botanical Congress, Berlin.

Goodall, Colin R. and Bookstein, F. L. (1987). Statistical aspects of biomedical imaging. *American Association for the Advancement of Science* annual meeting, Chicago, IL.

Goodall, Colin R., Kafadar, Karen, and Tukey, John W. (1990Ua). "An analysis of lung cancer mortality rates using urbanization and geography: Assessing sources of geographical variation," (prepared, in part, in connection with research at Princeton University sponsored by the Army Research Office (Durham) DAAL03-86-K-0073 and DAAL03-88-K-0045.)

Tukey, John W. (1989Ub). "Randomization and rerandomization: The wave of the past in the future," (presented to Ciminera Symposium, Philadelphia, June 1988; invited speaker at ETH-Zentrum, Zurich, Switzerland July 4, 1988).

Tukey, John W. (1989Uc). "Polyranges and natural approximation".

Tukey, John W. (1989Ud). "Introduction to modern analysis of variance," Murray Hill Statistics Seminar, AT&T Bell Laboratories, Murray Hill, NJ June 9, 1989.

Tukey, John W. (1989Ue) "The impact of the geophysical sciences on statistics and data analysis," S. S. Wilks Workshop, May 22-24, 1989.

Tukey, John W. (1989Uf). "The role of Statistics," (opening session speaker) American Statistical Association 150 Sesquicentennial meetings, Washington, D. C., August 6, 1989.

Tukey, John W. (1990Ug). A suggested, more unified approach to multiplicity".

Tukey, John W. (1990Uh). "Some set-ups for ANOVA-like inference". DAAL03-88-K-0045.)

3. Theses

Ph.D. Thesis

1986—

Nguyen, H., "Approximation of the optimum Pitman compromise estimate in O'Brien's case, investigated in terms of a single configuration," October.

1989—

Hansen, Katherine M., "Some statistical problems in geophysics and structural geology," June.

4. Technical Reports 1986—1990

Technical Reports: Department of Statistics, Princeton University

Number	Title	Author and date
291	Thinking about non-linear smoothers	John W. Tukey May 1986
292	Kinds of bootstraps and kinds of jackknives, discussed in terms of a year of weather-related data	John W. Tukey April 1987
293	Procedures for separations within batches of values, I. The orderly tool kit and some heuristics	Thu Hoang John W. Tukey March 1989
294	Tuning a major part of a clustering algorithm	Katherine M. Hansen John W. Tukey February 1988

296	(2nd DRAFT) Procedures for separations with batches of values, III. The case of unequal accuracy	Thu Hoang John W. Tukey July 1988
298	Procedures for separations within batches of values, II. More detailed heuristics and some simulation results	Thu Hoang John W. Tukey December 1988
299	Scrawl strips and letter or B-letter strips: depicting marginals of scatter plots	John W. Tukey James G. Veitch August 1989

Technical Reports, temporary series, Department of Statistics, Princeton University

Number	Title	Author and date
57	Principal components analysis of neural and facial skull configurations as a measurement of of orthocephalization in rats	C. R. Goodall A. Bose G. Das Gupta February 1986
58	Change-of-shape: a production system of S macros for growth analysis	C. R. Goodall March 1986
60	Characterization of skew-co-ordinate duality	C. R. Goodall May 1986

Technical Reports: Department of Civil Engineering, Princeton University

Number	Title	Author and date
SOR 87-9	Interpolation of multivariate data	Colin R. Goodall M. Thoma November 1987
SOR 87-11	The use of robust methods for shape comparisons	Colin R. Goodall
SOR 88-7	The analysis of averages and the analysis of variance	Colin R. Goodall April 1988

Internal Working Papers (IWP)

Number	Title	Author and date
IWP-71	Resistant fitting of quadratics to seven equally spaced points of which some may be missing	John W. Tukey 1987
IWP-72	Diagnostic tools for character clouds	John W. Tukey 1987
IWP-73	Comparing empirical distributions over time	John W. Tukey 1987
IWP-74	Some questions about categorical regression	John W. Tukey 1987
IWP-75	Three-word access with branching lengths	John W. Tukey 1987
IWP-81	Validating detectable differences	John W. Tukey 1988
IWP89-1	Percentage points of the range	John W. Tukey 1989
IWP89-2	Empirical improvement of HMM word-classing schemes	John W. Tukey 1989
IWP89-7	Introduction to paragrammar	John W. Tukey 1989
IWP89-8	Borrowing strength applied to 2-way tables of jackknifed variances	John W. Tukey 1989
IWP89-9	Higher criticism for individual significances in several tables or parts of tables	John W. Tukey 1989
IWP89-10	Class of accumulation patterns useful in support of certain agglomerative clustering algorithms	John W. Tukey 1989
IWP90-1	A collection of points relevant to making regression incisive	John W. Tukey 1990

Statistical software

Goodall, C. R. (1989) ANOVA: S functions for classical and resistant analysis of averages and the anlysis of variance using the sweep operator. S library archive, statlib@temper.stat.smu.edu

5. Sketch of work May 1986 — 1990

Over this period work was carried out on a considerable variety of topics, most directed toward improving the analysis of data. In the next 15 sections, this work is summarized and related to papers and reports under 15 headings: access, anova, clustering, graphical techniques, hints, MC, randomization, regression, robustness, shape, smoothing, stability, techniques (computational and statistical), and urbanization. Overviews and collected volumes are now discussed first.

The interface between data analysis and computing was reviewed briefly (Tukey 1986k). The impact of the geophysical sciences on statistics and data analysis has been considered (Tukey 1989Ue)

The likely evolution of the data analytic, and statistical, techniques likely to be used by the members of the American Statistical Association's Section on Physical and Engineering Sciences -- which will reflect quite well the techniques used across much broader areas of application - - was forecast - - and the relevance of describing the evolution of each of many data analytic techniques in terms of three consecutive 30-year periods was pointed out. (Tukey 1989w)

An update of "Future of Data Analysis (originally published by Tukey in (1962) was requested for an international meeting in France. A partial update, by Morgenthaler and Tukey (1990r), was prepared, presented, published. A fuller version is to be prepared, and is likely to be published in book form.

Volumes III to VI of Tukey's Collected Papers were issued, including 24 previously unpublished papers (see Publications, above).

6. Access

A variety of special topics related to access of full-text documents by searching full text have been explored (Internal Working Papers 75, 89-2, 89-7, 89-10).

7. Analysis of Variance (ANOVA)

Work on co-editing, and writing several chapters for, a book to be called *Fundamentals of Exploratory Analysis of Variance* continued during the large part of the period being reviewed. Appearance of the book is planned for Fall 1991 (co-editors: David C. Hoaglin, Frederick Mosteller, John W. Tukey, 1991h). This book takes a much more modern - - and much more realistic - - view of the analysis of variance than anything in print. At least one succeeding volume is in preparation.

Some significant innovations include:

- a more general - - more widely applicable - - basis for the "Rule of 2" in downsweeping combining some packets (lines) in an initial analysis with each other,
- serious thought about which comparisons in a 2-*way* table deserve special attention - - mainly "submaineffects" and "double differences" (2-way differences exhibiting interaction or its absence),
- use of biranges (= maximum size of double differences in a 2-way table) by analogy with ranges (= maximum size of differences in a 1-way table),
- simple approximations to birange % points,
- extensions to 3-way tables (only discussed lightly).

(More information about authored or coauthored chapters can be found under Publications, above). Some aspects of this work were summarized in (Tukey 1989Ud), others are alluded to in (Tukey 1991Uh).

The application of borrowing strength by median polish in the special case where the tabulated values are jackknived estimates of variance has been considered (IWP89-8, Tukey).

Other work will contribute to the second volume of this series including (Goodall and De Veau 1991U).

8. Clustering

The stage-by-stage development of a major portion of a clustering algorithm has been documented, submitted for publication, and revised (Hansen and Tukey 1990Sb, based on Technical Report 294). The resulting algorithm is quite novel combining a variety of quite distinct subalgorithms and, while it makes no explicit use of a Gaussian distributional assumption, it shows performance against a Gaussianity-distributed test bed that is almost as good as that provided by a Gausisan-likelihood-based algorithm. Thus its performance on real-world not-exactly-Gaussian data may well be better than any of the *many algorithms presently* available. It is notable that its development and evolution involve clusters that overlap one another seriously.

9. Graphical Techniques

Techniques for displaying distributions, mainly in terms of individual points of a sample have been explored, and innovative possibilities expounded (Tukey and Tukey 1990Se).

Diagnostic tools for examining character clouds have been discussed (IWP-72, Tukey).

Displaying linked aspects of data points has been reviewed and discussed (a technical report begun here will be reported under DAAL03-88-K-0045).

Delineation plots for bivariate data have been developed and discussed Goodall, Stoughton and Easton, 1986).

(Work after May 1, 1988 in this area is reported under DAAL03-88-K-0045.)

10. Hints

Exploratory data analysis can only serve its functions by detecting and mentioning phenomena, not all of which meet the usual standards (significance, confidence) of

confirmatory data analysis. *But mentioning anything and everything* dredged up in an extensive and deep exploration is equally unlikely to be helpful. Some guidance for the choice of what is to be mentioned is probably essential.

Catherine Marsh and John Tukey have been considering this problem for at least three years (since 1988) and draft discussions of what to use and how to use it are approaching readiness for publication. (An earlier version is Tukey 1990Ug).

11. Multiple Comparisons

A review paper on the Philosophy of Multiple Comparisons, originally a Miller lecture at Stanford (Tukey 1991n). This paper introduces - - and discusses - - a variety of issues of importance for mulitple comparisons. It interrelates substantially with the work on analysis of variance (see Section 7, above) and the work on hints (see Section 10).

An application of the "higher criticism" to the question - - what fractions of *individually significant results*, when all candidates are divided into bundles (perhaps one bundle for each of several tables), are likely to be real - - has been prepared (IWP89-9, Tukey).

The problem of approximating the distribution of the studentized birange (see Section 7, above) by a well-chosen studentized range distribution has been studied and discussed (Tukey 1989Uc).

12. Randomization

The state of the art of rerandomization as offering an almost completely trustworthy analysis of randomized experiments or data collections, as well as a comparatively highly trustworthy analysis of other data sets has been reviewed and extended (Tukey 1989Ub).

13. Regression (and related matters)

The problem of simple robust regression, in the face of both smoothly-varing variability and exotic values requiring robust estimation has been studied, and a substantial paper will appear (Cohen, Dalal, and Tukey 1991Sa).

The use of many covariates in analyzing timing-of-events experiments has been re-examined and new, effective techniques proposed (Tukey 1991o). Similar approaches should be effective in a wide variety of regression or regression-related circumstances.

Some questions about categorical regression have been considered (IWP74, Tukey).

14. Robustness

Earlier and continuing work on configural polysampling - - a realistic approach to optimum robustness - - has culminated in the appearance (Spring 1991) of a small book (Morgenthaler and Tukey 1991c).

Since a diverse set of techniques for robustly smoothing numerical sequences are now available, it is important to learn how to think about robust smoothers, and particularly about how to select a robust smoother for a particular purpose. These questions have been examined in some depth (Technical Report 291, Tukey). Some ways to make regression more incisive have been discussed (IWP90-1, Tukey).

The resistant fitting of straight lines to 9 or fewer points has been considered (IWP-71, Tukey)

15. Shape

Major emphases in this area include:

- the extension of Procrustes techniques, both least-square and robust, to the comparison of more than two (geometrical) forms (generalized Procrustes techniques)

- a statistical model for shape change by small perturbation
- placing the analysis of shape differences in a rigorous multivariate framework (F-tests)
- investigating descriptions of deformations by means of an ensemble of simple, low-order deformations of subregions
- a diversity of robustifications
- understanding the constraints on the residuals from a Procrustes fit
- extension of Procrustes fitting to weighted least squares
- inclusion of projective transformations in the hierarchy of transformations previously considered
- emphasis on geometric matching and image registration.

The first 3 of these points are reported in Bose and Goodall (1987). The others appear in other Goodall papers and reports in Section 3 and 4 above, or carry over into work under ARO DAAL03-88-K-0045, under which further work on this topic will be reported.

16. Smoothing

The robust smoothing of sequences, see Section 13 above (Technical Report 291).

Robust smoothing in the plane has received continuing attention. Important ideas include:

- an inverse convex-hull procedure for computing a polygon (or set of nested polygons) surrounding each data point,
- a convenient data structure for computing a median,
- adaptation of the end-value rule for boundary data.

Work on this topic continues.

17. Stability of results

The stability of adjusted (specifically age-adjusted) rates has been discussed (Tukey 1986u).

The use of resampling techniques - - jackknife or bootstrap - - to assess stability of results of data analysis in those situations where blocking is essential have been examined, and reasonable techniques for doing this have been discussed. (Technical Report 292, Tukey).

Techniques for deciding when it is desirable to discuss a batch of numerical values as two or more subbatches -- solely on the basis of the numerical values themselves -- have been examined. The first results are available as Technical Reports (293 and 298, Hoang and Tukey).

The g-and-h distributions form a useful 2-parameter family, accommodating skewness and elongation. Because they can be fitted in terms of quantiles (order statistics) they may prove considerably easier to estimate than families for which moment estimation seems natural (which turn out to decimal very large sample sizes). Empirical bounds for quantile-based estimates of g have been studied (Hoaglin and Tukey 1989Sc).

The degree to which the adequacy of an attempt to design an experiment of prescribed power can be assessed after the data has been collected has been discussed (IWP-81, Tukey).

The comparison of an ordered set of parallel distributions has been considered (IWP73, Tukey).

18. Techniques, computational

Interpolation in the plane, and in higher dimensions, has been studied, providing a common framework for data interpolation (related to key-frame interpolation) and view interpolation (related to kinematic displays of high dimensional data) (Goodall and Thoma

1987, and Technical Report SOR 87-9).

Functions useful in the analysis of averages have been coded and reported (Goodall, Technical Report SOR 88-7).

Software for the superimposition of forms has been prepeared.

Empirical formulas for unusual % points of the range have been prepared (IWP89-1, Tukey).

19. Techniques, statistical

A convenient, quite detailed table of the distribution of Student's t has been prepared and publsihed (Kafadar and Tukey 1988h).

Techniques for re-expressing exponentially distributed quantities, using simple hybrid re-expressions have been studied and will appear shortly (Tukey 1991p).

20. Urbanization measures

Novel but simple measures of urbanization for geographical units like counties ranging from:

- the logarithm of the size of the largest place

to

- the logarithm of the square root of the sum of the squares of the sizes of all places

have been tried out in such contexts as cancer rates, (age specific) birth rates, and median family incomes (Kafadar and Tukey 1990Sd, Goodall, Kafadar and Tukey 1990Ua).